Springboard--Intermediate Data Science Program
Capstone Project Proposal
By Intan Sari Septiana
December, 2019

Client and Background

Grab Holdings Inc, formerly known as MyTeksi and GrabTaxi, is a Singapore based transportation network company. In addition to transportation, the company offers food delivery and digital payments services via mobile app. Grab was originally founded in Malaysia and later moved its headquarters to Singapore. The company has been proactively pushing to make transportation in SEA safer. As part of the effort, they want to identify dangerous driving in a timely manner.

Problem Statement

Grab wants to identify dangerous driving by deriving a model with the telematics data for each trip and the label if the trip is tagged as dangerous driving. This project will focus on:

- Which variables are most relevant to predict dangerous driving?
- Since the data provided is time-series data, where one driving trip have multiple data points, which aggregation of telematics data is better (min, max, mean, etc)?

About the Dataset

The data came from the Grab AI For SEA website; (https://www.aiforsea.com/safety) with two files:

- Telematics data: The dataset used contains telematics data during trips (bookingID) provided by Grab. The telematics data is generated from a four-wheel driver's smartphone. Telematic data variables are:

Field	Description
bookingID	trip id
Accuracy	accuracy inferred by GPS in meters
Bearing	GPS bearing in degree
acceleration_x	accelerometer reading at x axis (m/s2)
acceleration_y	accelerometer reading at y axis (m/s2)
acceleration_z	accelerometer reading at z axis (m/s2)

gyro_x	gyroscope reading in x axis (rad/s)
gyro_y	gyroscope reading in y axis (rad/s)
gyro_z	gyroscope reading in z axis (rad/s)
second	time of the record by number of seconds
Speed	speed measured by GPS in m/s

- Label data: Each trip assigned with label 1 or 0 in a separate label file to indicate dangerous driving. Dangerous drivings are labelled per trip, while each trip could contain thousands of telematics data points.

Approach

For doing this project I plan to do these steps:

- Data Wrangling: since the data provided by the client is already in an easy format to process, it is anticipated that not much wrangling will be needed.
- Storytelling and Applications of Inferential Statistics: this phase involves activities such as generating summary statistics to see the data proportion and distribution; and find correlations with the target variable and look for interesting trends or patterns in the data with visualizations
- Baseline Modeling:Includes feature extraction from aggregating historical data and feature selection. It is anticipated that the baseline model will be Logistic Regression, commonly used for binary classification problems. I will use randomized search to find the best model parameters and use cross-validation to validate the model.
- Extended Modeling: after analyzing the performance of the baseline, other approaches will be investigated
- Evaluation: After building the baseline and other models a comparative analysis of their performance will be conducted with respect known metrics such as accuracy, precision, recall, etc. This analysis will result in the choosing of appropriate modeling approaches for the underlying business problem.

Deliverables

The deliverables of this project will be in the form of:

- Full report consists of descriptive analysis, modeling process and outcomes
- A presentation with the slide-deck
- Jupyter notebooks containing all codes from data exploration to evaluation